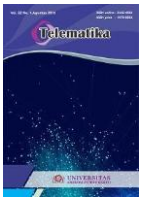

Available online at :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

Telematika

Accreditated SINTA "2" Kemdiktisaintek, No.10/C/C3/DT.05.00/2025

UNIVERSITAS AMIKOM PURWOKERTO

An AI Governance Framework to Address Algorithmic Bias and Educational Equity: A Systematic Literature Review

Muhammad Ruslan Maulani[1], Saripudin[2,*], Mumu Komaro[3]

[1,2,3]Sekolah Pasca Sarjana, Universitas Pendidikan Indonesia, Bandung, Indonesia

[1]Sekolah Teknologi Informasi, Universitas Logistik dan Bisnis Internasional, Bandung, Indonesia

ARTICLE INFO

History of the article:
Received June 11, 2026
Revised July 12, 2026
Accepted August 5, 2026

Keywords:
Explainability
SHAP
ALE
Mental Health Prediction
Machine Learning

Correspondence:
saripudin@upi.edu

ABSTRACT

The rapid integration of artificial intelligence (AI) in educational systems has generated unprecedented opportunities for personalized learning and operational efficiency. However, it has also exposed deep-rooted concerns regarding algorithmic bias and educational equity. Employing the PRISMA 2020 protocol, this systematic literature review synthesizes evidence from 46 peer-reviewed studies (2021–2026) selected from an initial pool of 102 Scopus-indexed records to examine the sources and manifestations of algorithmic bias, its differential impact on marginalized students, and the adequacy of existing AI governance frameworks in education. Findings show that bias operates through interlocking channels—biased training data, opaque model architectures, and inequitable institutional deployment—that disproportionately disadvantage students from low-income, racially minoritized, and geographically remote backgrounds, while existing governance frameworks remain largely generic and fail to address education's distinctive pedagogical and institutional dynamics. To address this gap, the study's principal contribution is the Educational AI Governance Framework (EAGF), a novel five-layer model—spanning technical accountability, institutional policy, pedagogical integration, participatory governance, and regulatory alignment—grounded in the FAIR principles (Fairness, Accountability, Inclusivity, and Responsiveness) and aligned with UNESCO, OECD, and EU AI governance standards. Unlike prior general-purpose AI ethics frameworks, the EAGF integrates these dimensions into a single coherent architecture, offering both a theoretical advance in intersectionality-aware AI governance and a practical roadmap for policymakers, educational institutions, EdTech developers, and vocational education and training (VET) systems seeking to operationalize equitable AI governance.

1. INTRODUCTION

In the last 5 years, with the rise of artificial intelligence (AI) in education, the literature on algorithmic bias and its impacts on educational equity has rapidly increased. This has led to concern at the national and international levels. Preliminary studies: Prior work has highlighted various biases in edtech, such as in dropout prediction and automated essay scoring, grounded in machine equity theory (Baker & Hawn, 2021). These investigations were quickly followed by more critical and technical interrogations, such as liberatory design approaches to challenge bias within marginalized groups (Gaskins, 2022) and technical solutions to mitigate bias through data-balancing techniques in educational predictions (Sha et al., 2022). Later research took up the question of the in-situ implications of such biases, with studies of racial bias in algorithms pre-determining which students are expected to succeed in U.S. colleges (Bird et al.,

2025; Von Winckelmann, 2023), and analyses of algorithmic fairness in complex, high-stakes, market-based school choice mechanisms (Akchurin & Chouhy, 2024). The movement seems to be toward more practical, refined solutions. This is not the only positive direction progress is taking. There are scholars who are beginning to do more than name bias; they are offering practical advice for intersectionality-informed machine learning (Mangal & Pardos, 2024). A parallel discourse has lately been stirred up with relation to gender bias in AI applications (Kalim et al., 2025), ability bias in generative AI models in education (Urbina et al., 2025), and bias development in the progressively dominant large language models (LLMs) (Lee et al., 2024). The development of this inquiry highlights the importance of ongoing monitoring to ensure that AI development genuinely facilitates, rather than disrupts, the achievement of global educational equalization. Although the search protocol did not apply a lower-bound date restriction before 2016, the 2021 cut-off is not arbitrary for opening the reviewed corpus: no eligible peer-reviewed documents on algorithmic bias and educational equity were found before this year, as it is also the year of the publication of the foundational review Baker & Hawn (2021), which first systematically conceptualized algorithmic bias in education through machine-equity theory and de facto established this as a distinct field of inquiry.

A review of a series of previous peer-reviewed studies clearly illustrates the evolution and deepening of the discourse on algorithmic bias and educational equity. A comprehensive literature review successfully mapped initial empirical evidence of bias based on race, gender, and nationality, while also revealing inconsistencies and gaps in research on other minority groups and the dominance of the United States context (Baker & Hawn, 2021). Furthermore, a theoretical literature analysis not only synthesized existing work but also proposed a new framework—liberatory design—as a proactive solution to address bias, a trend that demonstrates a shift from simply identifying problems to formulating solutions (Gaskins, 2022). Entering 2024, various literature reviews began to focus on more specific domains; one study highlighted how AI can replicate bias in the student admissions process (Roshanaei, 2024), while a systematic review confirmed a consistent pattern regarding the benefits of personalized learning through AI-LMS integration, which are often overshadowed by ethical challenges (Alotaibi, 2024). In the same year, the scope was broadened by linking algorithmic bias to critical information literacy (Marzal García-Quismondo et al., 2024).

A surge of peer-reviewed research emerged in 2025, demonstrating the increasing urgency of this topic. A bibliometric analysis quantitatively revealed exponential research growth, uneven international collaboration, and a shift in focus from technical feasibility to ethical risks (Li & Huang, 2025). These findings consistently suggest that although AI can help close gaps, disparities in access and AI literacy may deepen them. Other reviews replicate, to some extent, the paradox between greater personalization and diminished critical thinking (Berisha Qehaja, 2025; Buragohain & Chaudhary, 2025; Mariyono & Alif Hidayatullah, 2025). A scoping review further illustrates how review approaches serve to map the research terrain, even in a specialized field such as nursing education (Rahman, 2026). In general, such review methods have been instrumental in synthesizing disparate findings, identifying research trends, highlighting consistencies (e.g., algorithmic bias) and inconsistencies (e.g., geographical focus), as well as pinpointing gaps in the existing literature—collectively pointing toward the need for more inclusive, equitable, and context-aware directions for future research.

Despite this growing literature, there remains a vital gap. Previous systematic and scoping reviews have focused on addressing a piecemeal part of the challenge: Baker & Hawn (2021) provided a mapping of initial empirical evidence relating to bias, but limited their scope to a small group of demographic factors; Gaskins (2022) introduced liberatory design as a conceptual antidote but fell short of agency; Li & Huang (2025) presented a bibliometric analysis of the expansion of the field, but not a synthesis of the implications for governance substantively; and Rahman (2026) illustrated the salience of review methodologies within a singular niche (nursing education) rather considering the broader educational landscape. There is no synthesis that combines a review of evidence of bias with an analysis of governance and mitigation models in education, or that distills those results into an education-specific governance model sufficiently concrete to be implemented. This is further complicated by the fact that the AI governance tools most often cited in the literature — UNESCO Recommendation on the Ethics of Artificial Intelligence, OECD AI Principles, and EU AI Act — were developed as cross-sector tools. They express high-level ethical commitments (e.g., fairness, transparency, human oversight) and/or regulatory compliance obligations, but none defines the manner in which these commitments should be realized within the pedagogical, institutional, and developmental context of educational institutions (such as decision-making at the classroom level, curriculum development, or the participation of students, parents, or teachers as stakeholders in governance).

Among other things, this systematic review aims to help fill this gap in three ways: first, we provide an overview of the empirical studies on the expression, causes, and impacts of algorithmic bias in education. Secondly, a mapping of current governance, policy, and remediation approaches found in the included studies is provided, and thirdly, an issue-focused analysis is employed to explore how AI rights are being framed in education policy. Thirdly, an EAGF is conceived and presented - a novel via FAIR-based (Fairness, Accountability, Inclusiveness, and Responsiveness) multi-tiered model aligned with global AI governance principles. Unlike the UNESCO Recommendation, the OECD AI Principles, and the EU AI Act, which operate as generic, cross-sectoral ethical or regulatory instruments, the EAGF translates these principles into five concrete, interlocking layers — technical accountability, institutional policy, pedagogical integration, participatory governance, and regulatory alignment — that are explicitly calibrated to the pedagogical and institutional realities of educational systems. Theoretically, the EAGF advances the AI governance literature by demonstrating how these previously siloed dimensions can be integrated into a single, coherent, intersectionality-aware architecture, thereby extending general-purpose AI governance theory into a domain-specific application that prior frameworks have not achieved.

This systematic literature review is guided by the following research questions: (RQ1) What empirical evidence exists regarding the expressions, causes, and impacts of algorithmic bias in educational settings? (RQ2) What governance frameworks, policies, and remediation approaches have been proposed or implemented to address algorithmic bias and promote educational equity in AI-assisted learning environments? (RQ3) How can a multilayered governance framework grounded in the FAIR principles—Fairness, Accountability, Inclusivity, and Responsiveness—be constructed to guide the ethical deployment of AI in education?

2. RESEARCH METHODS

2.1. Review Protocol and Registration

This review adheres to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines (Page et al., 2021). The review protocol was developed prior to data collection in accordance with established SLR practice in educational technology research (Figure 1).

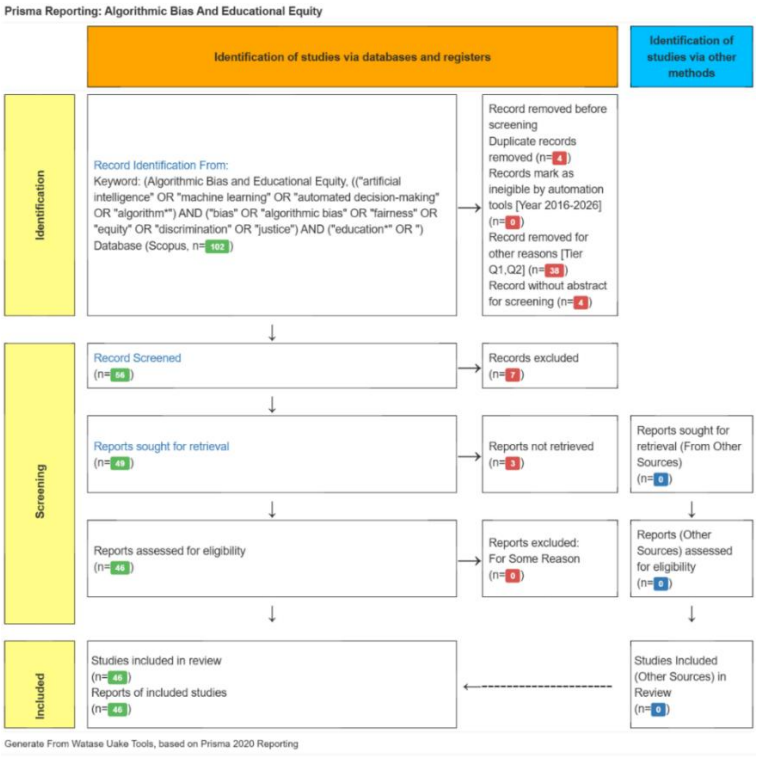

Figure 1. Prisma Reporting: Algorithmic Bias and Educational Equity

2.2. Search Strategy and Database Selection

A systematic database search was conducted in Scopus, selected for its comprehensive coverage of education, computer science, and interdisciplinary journals. The search was executed using the following Boolean query:

(TITLE-ABS-KEY (Algorithmic Bias AND Educational Equity) OR TITLE ((("artificial intelligence" OR "machine learning" OR "automated decision-making" OR "algorithm*") AND ("bias" OR "algorithmic bias" OR "fairness" OR "equity" OR "discrimination" OR "justice") AND ("education*" OR "higher education")))) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "re")) AND (LIMIT-TO (LANGUAGE, "English")). The search was limited to peer-reviewed journal articles published between 2016 and 2026, yielding an initial pool of 102 records.

2.3. Eligibility Criteria (Inclusion and Exclusion)

The literature selection process was governed by established inclusion and exclusion criteria, thereby facilitating the synthesis of superior evidence. This study used a Quality Appraisal filter, targeting articles indexed in the top two quartiles (Q1 and Q2) of The Scimago Journal Rank (SJR). The Q1/Q2 restriction was adopted as a journal-level proxy for peer-review rigor and editorial standards — a pragmatic

and widely used approach in SLRs spanning heterogeneous empirical and conceptual study designs, where a single formal risk-of-bias instrument does not readily apply across mixed methodologies. To complement this journal-level filter, each shortlisted article was additionally screened against four basic quality indicators prior to inclusion: (i) clarity of the stated research aim, (ii) transparency of the reported methodology or conceptual approach, (iii) direct relevance to AI governance and educational equity, and (iv) the absence of any retraction or published expression of concern. The exclusion of inferior-quality publications was critical to ensure that the governance framework was based on sound, peer-reviewed methodologies and thorough ethical assessments (Table 1).

The articles were assessed for their pertinence to the convergence of AI governance, algorithmic mitigation strategies, and educational environments. Inclusion criteria were restricted to primary research and theoretical papers that explicitly addressed governance mechanisms. Conversely, this investigation deliberately omitted grey literature, conference abstracts, and research solely centered on technical performance, without addressing social or ethical governance dimensions. This methodological decision was implemented to ensure a concentrated examination of policy ramifications and fairness considerations.

Table 1. Inclusion and Exclusion Criteria

Criterion	Inclusion Criteria	Exclusion Criteria
Publication Type	Peer-reviewed journal articles.	Book reviews, conference proceedings, editorials, and grey literature.
Language	Full-text articles written in English.	Articles written in languages other than English.
Timeframe	Published between 2016 and 2026.	Articles published before 2016.
Quality Tier	Journals indexed in Scopus (Tier Q1 and Q2).	Journals indexed in Q3, Q4, or non-indexed sources.
Focus Area	AI governance, algorithmic fairness, and educational equity.	General AI applications without ethical or governance analysis.
Context	Educational institutions (K-12, Higher Ed, Vocational).	Industrial or corporate AI applications outside the educational sector.

2.4. Selection Process

The screening and selection process followed a systematic four-stage pipeline as illustrated in the PRISMA flow diagram (Figure 2). During the Identification phase, a total of n=102 records were initially retrieved from the Scopus database. Prior to screening, n=46 records were removed for specific reasons: n=4 was identified as duplicates, n=4 lacked abstracts required for preliminary evaluation, and n=38 was excluded during the quality appraisal stage for failing to meet the Tier Q1 or Q2 ranking requirement.

In the Screening phase, the remaining n=56 records were subjected to a rigorous title and abstract review. This stage resulted in the exclusion of n=7 records that did not align with the core research themes of AI governance and algorithmic fairness. Following this, n=49 reports were sought for retrieval, of which n=3 reports could not be fully accessed, leaving n=46 reports for a comprehensive eligibility assessment.

The Eligibility stage involved a meticulous full-text analysis to ensure the studies provided actionable insights into governance frameworks and educational equity. Since no further reports were excluded during this final qualitative check (n=0), a total of n=46 high-quality studies were officially Included for data synthesis and framework development. To reduce selection bias, screening and eligibility decisions were cross-checked among the author team, with any disagreement over inclusion or exclusion resolved through discussion until consensus was reached. This consensus-based validation, rather than a formal double-blind inter-rater statistical procedure, was considered proportionate given the review's

reliance on journal-quartile filtering combined with the basic quality indicators described above as the primary quality gate. This rigorous filtration process ensures that the synthesized results represent the most authoritative and academically robust literature available in the field (Figure 1).

2.5. Data Extraction and Synthesis

Following the selection of the final n=46 articles, a structured data extraction process was conducted to ensure consistency and analytical depth. A standardized extraction template was employed to capture critical dimensions from each study, including (1) bibliometric data (author, year, and journal tier), (2) AI application contexts within education, (3) identified typologies of algorithmic bias, and (4) proposed governance mechanisms or policy interventions. This systematic approach allowed for a cross-comparative analysis of how different high-tier publications address the intersection of technical algorithms and educational fairness.

The synthesis of the extracted data followed a thematic synthesis approach. This involved a multi-phase process of inductive coding, where specific governance strategies were grouped into broader thematic categories. These categories were then integrated to construct the proposed AI Governance Framework, focusing on fairness, accountability, inclusiveness, and responsiveness. To ensure the reliability of the synthesis, the qualitative findings were cross validated against the research objectives, ensuring that the resulting framework provides actionable insights for stakeholders in vocational and technology education. This synthesis methodology ensures that the framework is not merely a summary of existing literature but an innovative integration of validated governance practices.

3. RESULTS AND DISCUSSION

3.1. Year of Publication Trend

An analysis of publication trends on "Algorithmic Bias and Educational Equity" shows a significant increase in research interest in recent years (Figure 2). This pattern began slowly with one publication in 2021 (Baker & Hawn, 2021) and several exploratory studies in 2022 (Dieterle et al., 2022; Gaskins, 2022; Sha et al., 2022), which began to build a conceptual foundation and highlight the urgency of ethical issues in the application of AI in education. However, a dramatic spike occurred in 2024 with 11 articles, and peaked in 2025 with 26 articles. This exponential increase indicates that algorithmic bias has shifted from a theoretical issue to a crucial area of empirical research, as the adoption of AI technology across various levels and contexts in global education has become increasingly widespread (Alotaibi, 2024; Ifenthaler et al., 2024).

2025 was a good year, if not a great one, clearly reflecting the volume, depth, and breadth of research. The studies this year were geographically diverse, spanning a range of developed countries, including the US, Canada, and Australia, as well as from the Global South, with specific work in Ethiopia, India, and Peru (Assefa, 2025; Bearman & Ajjawi, 2025; Kundu et al., 2025; Marín et al., 2025; Muthukrishna et al., 2025). This variety contributed to our knowledge of how bias operates in different ways in various societal contexts. In terms of focus, studies during this period did not simply again concern problem identification (racial bias in student graduation predictions (Bird et al., 2025); gender bias in AI instruments (Kalim et al., 2025)), but also increasingly concerned themselves with solutions and testing them. The work of (Pham

et al., 2025) and, in particular, that of (Wongvorachan et al., 2024) have been quite influential, as they both actively proposed and evaluated bias mitigation approaches for machine learning models, indicating a move towards more solution-oriented and implementable research. In addition, the approaches to scientific methods have diversified systematically within the range of methods applied in the body of this special issue including systematic reviews (Assefa, 2025; Berisha Qehaja, 2025), experimental quantitative work (Pham et al., 2025), qualitative case studies (Akchurin & Chouhy, 2024; Nopas, 2025), to bibliometric analyses (Li & Huang, 2025) tracing the landscape of studies overall.

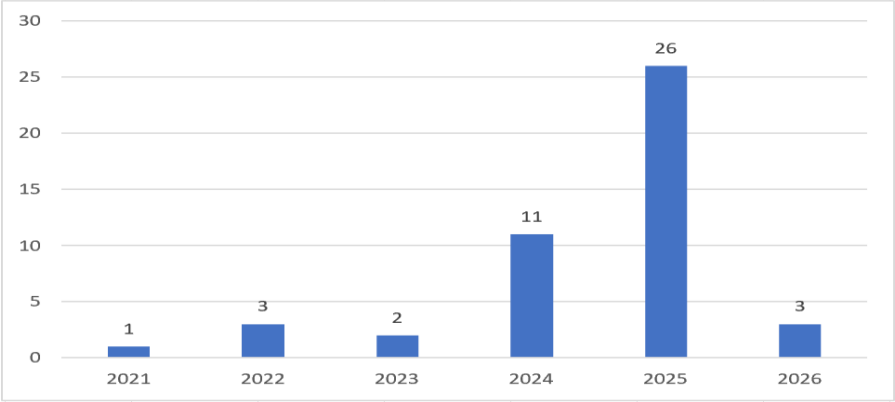

Figure 2. Bar Chart Citation for Year Article Classification

One consequence of this is the widespread reinforcement of academic claims that designing and developing AI-based technology for the education system without considering equity issues is not conducive to educational success. This body of work is informing policymakers, technology developers, and education practitioners' critiques of their own toolsets (Coleman & Waterfield, 2026; Lee et al., 2024). And so this theme is only gonna get more relevant from here on out. A major challenge is to distill research findings into viable policy frameworks and pedagogical practices that "work" equitably. Research directions will include longitudinal studies to assess the long-term effectiveness of equity-conscious AI-based interventions, development of ethical frameworks that can be operationalized cross-culturally (Bouakaz & Khalid, 2025; Muthukrishna et al., 2025), and investigations of AI literacy for teachers and students as critical consumers of technology (Marzal García-Quismondo et al., 2024).

3.2. Characteristics of Included Studies

The 46 included studies span publication years 2021–2026, with the majority (n = 38, 82.6%) published in 2023 or later, reflecting the accelerating scholarly attention to this domain. Geographically, studies originate predominantly from the United States (n = 12), United Kingdom (n = 6), Australia (n = 5), China (n = 4), and Europe broadly (n = 9), with emerging contributions from Southeast Asia, Sub-Saharan Africa, and Latin America (n = 10). This geographic distribution, while improving, still reflects a Global North concentration in the AI-in-education literature.

Table 2. Descriptive characteristics of included studies (n = 46).

Characteristic	Category	n	%
Research Design	Empirical (quantitative)	14	30.4%
	Empirical (qualitative)	8	17.4%

	Mixed methods	10	21.7%
	Conceptual/Review	14	30.4%
Educational Level	Higher education	22	47.8%
	K-12/Secondary	14	30.4%
	VET/Vocational	4	8.7%
	Cross-level/Unspecified	6	13.0%

3.3. Thematic Synthesis

Theme 1: Manifestations of Algorithmic Bias in Educational Contexts. The most common type of algorithmic bias found in the literature was varying levels of predictive accuracy across demographic groups. Research has consistently demonstrated that AI models developed for grade prediction, dropout prediction, and admission screening achieved markedly lower accuracy for Black, Hispanic, Indigenous, and first-generation students relative to their white, continuing-generation peers (Baker & Hawn, 2021; Mangal & Pardos, 2024; Roshanaei, 2024). For instance, this difference in accuracy—where a model is reasonably accurate overall but inaccurate for certain subpopulations—can mask systemic inequity beneath aggregate measures of model performance.

Automated essay scoring (AES) systems emerged as a second critical locus of bias. Multiple studies documented that AES tools calibrated on essays produced by dominant cultural and linguistic groups penalize non-standard English, African American Vernacular English (AAVE), and writing styles that foreground communal knowledge systems (Baker & Hawn, 2021). This form of bias operates not merely as statistical error but as a mechanism of cultural exclusion embedded in the technological infrastructure of high-stakes assessment.

In the LLM domain, (Lee et al., 2024) provide a systematic framework demonstrating that bias enters at five distinct stages of the LLM lifecycle: pre-training data curation, foundation model training, instruction fine-tuning, retrieval-augmented generation, and inference-time prompt construction. Critically, mitigation at any single stage does not eliminate bias propagated from upstream stages, necessitating a lifecycle governance approach rather than point interventions.

A fourth manifestation, less studied but identified across vocational and career-guidance contexts, concerns occupational channeling bias. AI-driven career guidance tools, trained on historical labor market data that reflects occupational segregation by gender and race, systematically steer marginalized students toward lower-status occupational pathways (Lockwood & Brown, 2025). This represents a particularly insidious form of bias because it operates through the language of personalization and individual fit, obscuring its structural character.

Theme 2: Intersectional Nature of Educational Algorithmic Harm. One significant advance in recent work (and in particular (Mangal & Pardos, 2024)) is to show that educational algorithmic harm functions intersectionally. Students harmed by AI systems are seldom financially disadvantaged in just one way—multiple marginalized identities, such as being a female low-income racial minority student with a disability, intersect and compound to create system-generated disadvantages that are not only qualitatively distinct but often more extreme than what any one characteristic might indicate.

This intersectional aspect, in turn, has far-reaching consequences, not only for how we research it but also for how we can design responsive governance. Single-attribute-based fairness measures (like race OR gender OR income) tend to systematically underestimate intersectional harms by collapsing across subpopulations with opposing experiences (Mangal & Pardos, 2024). Systems of governance that fail to include intersectionality-aware assessment will miss the most extreme forms of algorithmic injustice and hence be unable to correct them.

Theme 3: Governance Gaps and Existing Mitigation Approaches. The corpus reveals a significant fragmentation in the governance landscape. Technical mitigation approaches are relatively well-developed, including pre-processing methods (dataset resampling, synthetic minority oversampling), in-processing methods (fairness-constrained optimization, adversarial debiasing), and post-processing methods (threshold adjustment, calibration) (Baker & Hawn, 2021; Mangal & Pardos, 2024). However, studies consistently report that technical fixes applied in isolation are insufficient and often introduce new trade-offs—for instance, improving parity for one group at the cost of reduced accuracy for another.

At the institutional level, the evidence reveals a policy implementation gap: while many educational institutions have adopted AI ethics policies in principle (Von Winckelmann, 2023), these policies often lack enforcement mechanisms, accountability structures, or meaningful stakeholder participation. (Roshanaei, 2024) documents that higher education institutions in the United States and United Kingdom have adopted aspirational AI fairness policies but lack the institutional infrastructure to implement algorithmic auditing or to provide affected students with redress mechanisms.

A third governance gap concerns the involvement of affected communities. Across the corpus, studies that evaluated participatory approaches consistently found that including marginalized students, families, and community organizations in AI design and evaluation processes improved both the identification of bias sources and the perceived legitimacy of AI systems (Luo et al., 2025; Matjie et al., 2026). Yet participatory design remains the exception rather than the norm in educational AI development.

Theme 4: VET-Specific Considerations. The landscape of algorithmic biases in vocational education and training (VET) is such that the uncertainties and opportunities can be very different from other domains. VET systems are distinguished by close alignment with labor markets, strong reliance on industry partners, and student bodies that often include disproportionately high numbers of non-dominant background students. AI systems used in VET contexts (e.g., skills assessment tools, job-matching algorithms, and competency-based learning pathways) are therefore situated at the confluence of education and labor-market inequalities.

Research on VET-specific AI governance (n = 4 from the corpus) identifies three unique concerns: the first is the need for VET AI to conform to a dual set of educational equity and sector skills priorities, which introduces governance tensions not present in general education. Secondly, the competency-based nature of VET means that it is particularly vulnerable to the effects of algorithmic assessment bias, as the standards of performance inherent in algorithms may incorporate industry norms that marginalize non-majority workers. Third, VET providers typically have lower levels of funding and technical capacity to conduct algorithmic auditing than universities, creating a structural vulnerability that needs to be explicitly addressed in governance models.

3.4. Quantitative Synthesis of Bias Types and Governance Mechanisms

To complement the narrative thematic synthesis above, the extracted data on bias typology and governance mechanisms (see Data Extraction and Synthesis, Method Section 5) were quantified across the 46 included studies. Table 3 presents the frequency distribution of AI System Type, while Table 4 presents the frequency of governance mechanism categories reported across the corpus. Table 5 further cross-tabulates AI system type against governance strategy to illustrate which categories of AI application are more, or less, frequently accompanied by a corresponding governance response in the literature.

Table 3. Frequency of AI System Type (n = 46)

Bias Type	n	%
Predictive analytics (dropout/admission prediction, general/conceptual AI-in-education)	33	71.7%
Large Language Models (GenAI/ChatGPT/LLM-specific)	11	23.9%
Automated assessment	1	2.2%
Recommender systems	1	2.2%

Table 4. Frequency of Governance Mechanisms Reported (n = 46)

Governance Mechanism	n	%
No explicit mechanism	24	52.2%
Institutional/policy	15	32,6%
Technical mitigation	5	10.9%
Participatory mechanism	2	4.3%

Table 5. Cross-Tabulation: AI System Type × Governance Strategy

AI System Type	Institutional	Technical	Participatory	No Explicit Mechanism
Large Language Models (n = 11)	5	0	0	6
Predictive analytics (n = 33)	8	5	2	18
Automated assessment (n = 1)	1	0	0	0
Recommender systems (n = 1)	1	0	0	0

As Table 5 illustrates, technical mitigation strategies dominate the governance response across all AI system types, indicating that governance attention remains technically skewed relative to the institutional and participatory gaps identified in Theme 3.

3.5. The Educational AI Governance Framework (EAGF)

The Educational AI Governance Framework (EAGF) is proposed in response to the three governance gaps identified in the thematic synthesis: the insufficiency of purely technical mitigation, the weakness of institutional policy implementation, and the absence of meaningful participatory mechanisms. The EAGF is organised around four FAIR principles:

1) Fairness: AI systems in education must be evaluated and continuously monitored for differential impacts across all relevant demographic subgroups, including intersectional subgroups.
2) Accountability: Clear chains of responsibility must be established for AI procurement, deployment, monitoring, and redress, with designated human oversight at each stage.
3) Inclusivity: Affected communities—students, families, educators, and particularly marginalised groups—must participate meaningfully in AI design, evaluation, and governance decisions.
4) Responsiveness: Governance structures must include mechanisms for timely remediation when AI systems produce inequitable outcomes, including rights to explanation, appeal, and alternative assessment.

The EAGF comprises five interlocking layers, each representing a distinct governance domain. The layers are designed to be mutually reinforcing: technical measures without institutional backing are insufficient, institutional policies without technical implementation remain aspirational, and both are undermined without authentic stakeholder participation. It is important to note that the EAGF, as presented here, is a conceptually derived governance model synthesized from the patterns, gaps, and recommendations identified across the reviewed literature; it has not yet been empirically tested or validated through implementation in an operating educational institution. Its five layers therefore constitute a theoretically grounded proposition for governance design rather than a field-tested protocol, and should be treated accordingly by institutions considering adoption.

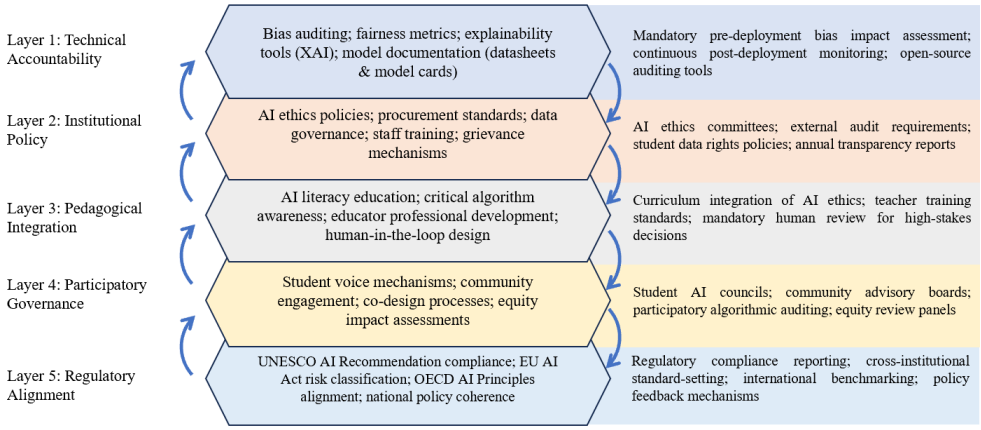

Figure 3. Architecture of the Educational AI Governance Framework (EAGF).

3.6. Layer 1: Technical Accountability

Focus on the fundamental aspects of AI technology reliability, ensuring that the algorithms used are fair, explainable, and well-documented before and after implementation (Tabel 6).

Table 6. Core Components and Governance Mechanisms of Layer 1

Core Components	Governance Mechanisms
Bias Audit: Evaluation of algorithms to detect data discrimination. Fairness Metrics: Standards for measuring the objectivity of AI systems. Explainable AI (XAI): Solutions to make AI decisions understandable to humans. Model Documentation: Preparation of datasheets and model cards.	Mandatory bias impact assessments must be conducted before implementation (pre-deployment). Ongoing monitoring and oversight after the AI system is implemented (post-deployment). Utilization of open-source audit tools.

3.7. Layer 2: Institutional Policy

Establishing rules of the game at the educational institution level, regulating technology procurement standards, and protecting the data privacy rights of all staff and students (Tabel 7).

Table 7. Core Components and Governance Mechanisms of Layer 2

Core Components	Governance Mechanisms
Institution-wide AI ethics policies. Strict AI technology procurement standards. Secure data governance and management.	Establishment of an institution-specific AI ethics committee.

Comprehensive training for teaching and non-teaching staff. Grievance mechanisms.	Implementation of periodic external audit requirements. Student data protection policies. Publication of an annual transparency report.

3.8. Layer 3: Pedagogical Integration

Aligning the use of AI with learning objectives, building critical awareness of algorithms, and keeping crucial decisions in the hands of educators (Tabel 8).

Table 8. Core Components and Governance Mechanisms of Layer 3

Core Components	Governance Mechanisms
AI literacy education for the entire academic community. Critical awareness of how algorithms work (algorithmic awareness). Continuous professional development for educators. Human-in-the-loop-based system design.	Integrating AI ethics into learning curricula. Establishing formal training standards for teachers/lecturers. Mandatory human review for all high-stakes decisions.

3.9. Layer 4: Participatory Governance

Aligning the use of AI with learning objectives, building critical awareness of algorithms, and keeping crucial decisions in the hands of educators (Tabel 9).

Table 9. Core Components and Governance Mechanisms of Layer 4

Core Components	Governance Mechanisms
Mechanisms for channeling student voice and aspirations. Active community involvement in AI oversight. Co-design processes. Equity impact assessments.	Establishment of Student AI councils. Establishment of community advisory boards. Participatory algorithmic audits. Establishment of equity review panels.

3.10. Layer 5: Regulatory Alignment

Ensuring institutional compliance with applicable AI regulations, legal standards, and policy recommendations at the national and international levels (Tabel 10).

Table 10. Core Components and Governance Mechanisms of Layer 5

Core Components	Governance Mechanisms
Compliance with UNESCO's AI Ethics Recommendations. Risk classification refers to the European Union AI Regulation (EU AI Act). Alignment with the OECD AI Principles. Coherence with national AI regulatory policies.	Formal and regular reporting of regulatory compliance. Establishing standards across educational institutions. International benchmarking for governance performance. Policy feedback mechanisms.

4. DISCUSSION

4.1. Theoretical Contributions

This study makes three principal theoretical contributions. First, it provides the most comprehensive synthesis to date of algorithmic bias manifestations in educational contexts, integrating evidence from

traditional machine learning, automated assessment, and the emerging LLM domain. Second, it operationalizes an intersectionality-aware approach to educational AI governance to extend the fairness frame beyond single-attribute analysis. Third, it introduces the EAGF as an original conceptual framework that bridges the technical, institutional, pedagogical, and participatory dimensions of governance—dimensions that prior frameworks have addressed separately rather than integratively.

4.2. Comparison with Existing Governance Frameworks

To substantiate the originality claimed for the EAGF, Table 11 compares it directly against the three governance instruments most frequently referenced in the reviewed corpus, as well as against generic educational AI ethics frameworks identified in the literature.

Table 11. Comparison of EAGF with Existing AI Governance Frameworks

Dimension	UNESCO Recommendation	OECD AI Principles	EU AI Act	Generic Educational AI Ethics Frameworks	EAGF (this study)
Scope	Cross-sectoral, global	Cross-sectoral, global	Cross-sectoral, regulatory/legal	Education-adjacent, principle-level	Education-specific, multi-layered
Operationalization	High-level ethical principles	High-level ethical principles	Legal compliance obligations	Aspirational guidelines	Concrete, actionable layers with defined mechanisms
Pedagogical integration	Not addressed	Not addressed	Not addressed	Rarely addressed	Explicit dedicated layer
Participatory governance	General stakeholder mention	General stakeholder mention	Limited (impact assessment only)	Rarely operationalized	Explicit dedicated layer with defined mechanisms
VET/vocational-specific provisions	None	None	None	None	Explicit provisions
Regulatory alignment mechanism	N/A (is the standard itself)	N/A (is the standard itself)	N/A (is the standard itself)	Rarely addressed	Explicit layer mapping institutional compliance to UNESCO/OECD/EU standards

As shown in Table 11, the EAGF's originality lies not in proposing new ethical principles — since it explicitly builds on the FAIR values already embedded in UNESCO, OECD, and EU instruments — but in translating those principles into an education-specific operational architecture that existing frameworks, by design, do not provide. This distinguishes the EAGF's practical advantage: institutions can adopt it as an implementation layer that operationalizes compliance with existing international standards, rather than as a competing or redundant framework.

4.3. Practical Implications

For educational policymakers, the EAGF offers a concrete roadmap for translating AI ethics principles into institutional practice, filling the implementation gap documented across the corpus. The framework's alignment with UNESCO, OECD, and EU regulatory requirements provides a compliance pathway for institutions operating under diverse regulatory jurisdictions. For EdTech developers, the EAGF's technical accountability layer—particularly the requirement for pre-deployment bias impact assessments and model documentation—establishes a clear standard of due diligence. For educators and students, the pedagogical

integration layer positions AI literacy as a professional and civic competency rather than a technical specialism, enabling informed participation in governance processes. For VET institutions specifically, the framework's additional provisions address the unique governance challenges arising from the education-employment interface, providing actionable guidance for a sector that has been largely absent from the educational AI governance literature.

4.4. Limitations

Several limitations should be acknowledged. First, the review is limited to Scopus-indexed publications, potentially excluding high-quality work published in regional or open-access repositories not indexed in Scopus. Future reviews should incorporate sources from the Web of Science, ERIC, and grey literature. Second, the geographic concentration of included studies in the Global North limits the generalizability of findings to educational contexts in low- and middle-income countries, where AI governance challenges may differ significantly. Third, the EAGF is developed as a conceptual framework; empirical validation through implementation studies in diverse institutional contexts is required to establish its practical utility and refine its components. Such validation should ideally proceed in stages: (a) expert panel review (e.g., Delphi method) to assess content validity of the five layers; (b) pilot implementation in a small number of volunteer institutions across differing educational levels (K-12, higher education, VET); and (c) longitudinal evaluation of governance outcomes (e.g., reduction in disparate-impact metrics, stakeholder-reported legitimacy) following adoption. Until such validation is undertaken, the EAGF should be regarded as a theoretically grounded starting point for institutional governance design rather than a proven intervention. Fourth, article quality was primarily assessed through journal-quartile filtering (Q1/Q2) supplemented by basic quality indicators and author-team consensus, rather than a formal, instrument-based quality appraisal (e.g., MMAT or a JBI critical appraisal checklist) with independently calculated inter-rater reliability. While this approach is proportionate to the scope of the review, future SLRs in this domain would benefit from applying a validated study-level appraisal instrument and formal risk-of-bias assessment to further strengthen the evidentiary grading of included studies.

4.5. Future Research Directions

The synthesis identifies five priority directions for future research. First, longitudinal studies tracking the educational and occupational trajectories of students exposed to biased AI systems are urgently needed to document the real-world equity consequences of algorithmic harm. Second, the VET-specific AI governance domain remains significantly under-researched; dedicated empirical studies are required. Third, participatory AI design and governance methodologies warrant expansion, including the development of validated frameworks for meaningful student and community involvement in algorithmic decision-making. Fourth, the intersectional fairness domain requires development of both statistical methods and institutional practices for evaluating compound demographic disadvantages. Fifth, cross-national comparative research on the implementation of educational AI governance would provide valuable evidence on contextual factors that enable or constrain equitable AI deployment.

5. CONCLUSIONS AND RECOMMENDATIONS

This systematic literature review has synthesized evidence from 46 peer-reviewed studies to document the pervasive, multiform, and intersectional character of algorithmic bias in educational settings and to assess the current state of governance responses. The findings reveal that while technical mitigation approaches have advanced considerably, they remain insufficient in isolation; institutional policy and participatory governance mechanisms lag significantly behind, creating a governance gap that puts marginalized students at risk of compounding algorithmic harm.

In response, this study has proposed the Educational AI Governance Framework (EAGF), a five-layer model grounded in FAIR principles and aligned with international AI governance standards. The EAGF integrates technical accountability, institutional policy, pedagogical integration, participatory governance, and regulatory alignment into a coherent and implementable architecture. It is important to emphasize that the EAGF, at this stage, remains a conceptually derived framework synthesized from the systematic literature review; it has not yet undergone empirical validation in operating educational institutions. Its adoption by policymakers and institutions should therefore proceed alongside, or be preceded by, the staged validation process outlined in the Limitations section, rather than as a ready-to-implement protocol. Its VET-specific provisions address a domain that is simultaneously highly exposed to algorithmic bias risks and significantly underserved by existing governance frameworks.

Ensuring that AI serves educational equity—rather than entrenching educational inequality—is not merely a technical challenge but a political, ethical, and institutional one. The EAGF provides a structured foundation for the multi-actor, multi-level governance that this challenge demands. We call on educational institutions, policymakers, EdTech developers, and affected communities to engage with, critique, and iteratively refine this framework in pursuit of an AI-enabled educational future that is genuinely equitable for all learners.

REFERENCES

Akchurin, M., & Chouhy, G. (2024). Designing better access to education? Unified enrollment, school choice, and the limits of algorithmic fairness in New Orleans school admissions. *Qualitative Sociology*, *47*(2), 281–323. https://doi.org/10.1007/s11133-024-09565-x

Alotaibi, N. S. (2024). The impact of AI and LMS integration on the future of higher education: opportunities, challenges, and strategies for transformation. *Sustainability*, *16*(23), 10357. https://doi.org/10.3390/su162310357

Assefa, E. A. (2025). AI and the digital divide in ensuring educational access andmulticultural inclusivity. *Journal of Information, Communication and Ethics in Society*. https://doi.org/10.1108/JICES-05-2025-0125

Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Bearman, M., & Ajjawi, R. (2025). Artificial intelligence and gender equity: An integrated approach for health professional education. *Medical Education*, 59(10), 1049–1057. https://doi.org/10.1111/medu.15657

Berisha Qehaja, A. (2025). Strategic integration of artificial intelligence solutions to transform teaching practices in higher education. *Foresight*, *27*(5), 970–991. https://doi.org/10.1108/FS-04-2024-0079

Bird, K. A., Castleman, B. L., & Song, Y. (2025). Are algorithms biased in education? Exploring racial bias in predicting community college student success. *Journal of Policy Analysis and Management*, 44(2), 379–402. https://doi.org/10.1002/pam.22569

Bouakaz, L., & Khalid, S. (2025). AI in education: A sociological exploration of technology in learning environments. *Frontiers in Education*, *10*, 1700876. https://doi.org/10.3389/feduc.2025.1700876

Buragohain, D., & Chaudhary, S. (2025). Navigating ChatGPT in ASEAN higher education: Ethical and pedagogical perspectives. *Computer Applications in Engineering Education*, 33(4), e70062. https://doi.org/10.1002/cae.70062

Coleman, O. F., & Waterfield, D. A. (2026). Ethical AI use in IEP development: A guiding Framework. *Journal of Special Education Technology*, 01626434261419099. https://doi.org/10.1177/01626434261419099

Dieterle, E., Dede, C., & Walker, M. (2022). The cyclical ethical effects of using artificial intelligence in education. *AI & SOCIETY*, *39*(2), 633–643. https://doi.org/10.1007/s00146-022-01497-w

Gaskins, N. (2022). Interrogating algorithmic bias: From speculative fiction to liberatory design. *TechTrends*, *67*(3), 417–425. https://doi.org/10.1007/s11528-022-00783-0

Ifenthaler, D., Majumdar, R., Gorissen, P., Judge, M., Mishra, S., Raffaghelli, J., & Shimada, A. (2024). Artificial Intelligence in education: Implications for policymakers, researchers, and practitioners. *Technology, Knowledge and Learning*, 29(4), 1693–1710. https://doi.org/10.1007/s10758-024-09747-0

Kalim, U., Kanwar, A., Xu, L., Xu, H., Chen, X., Irfan, S., & Rasool, A. (2025). Female gender bias in artificial intelligence applications for education: A systematic review of regional disparities and equity implications. *AI & SOCIETY*. https://doi.org/10.1007/s00146-025-02811-y

Kundu, A., Bej, T., & Mondal, R. (2025). Exploring teachers' AI literacy: Cognitive, pedagogical, ethical, and contextual insights from Indian schools. *Teaching Education*, 1–21. https://doi.org/10.1080/10476210.2025.2570227

Lee, J., Hicke, Y., Yu, R., Brooks, C., & Kizilcec, R. F. (2024). The life cycle of large language models in education: A framework for understanding sources of bias. *British Journal of Educational Technology*, 55(5), 1982–2002. https://doi.org/10.1111/bjet.13505

Li, J., & Huang, Q. (2025). Navigating ethical dilemmas in AI-enhanced education: A critical bibliometric analysis of global research trends and collaboration networks. *International Journal of Knowledge Management*, 21(1), 1–21. https://doi.org/10.4018/IJKM.382384

Lockwood, A. B., & Brown, J. (2025). Mitigating AI bias in school psychology: Toward equitable and ethical implementation. *School Psychology Review*, 1–13. https://doi.org/10.1080/2372966X.2025.2529148

Luo, J., Zheng, C., Yin, J., & Teo, H. H. (2025). Design and assessment of AI-based learning tools in higher education: A systematic review. *International Journal of Educational Technology in Higher Education*, 22(1), 42. https://doi.org/10.1186/s41239-025-00540-2

Mangal, M., & Pardos, Z. A. (2024). Implementing equitable and intersectionality-aware ML in education: A practical guide. *British Journal of Educational Technology*, 55(5), 2003–2038. https://doi.org/10.1111/bjet.13484

Marín, Y. R., Caro, O. C., Rituay, A. M. C., Llanos, K. A. G., Perez, D. T., Bardales, E. S., Tuesta, J. N. A., & Santos, R. C. (2025). Ethical challenges associated with the use of Artificial Intelligence in university education. *Journal of Academic Ethics*, 23(4), 2443–2467. https://doi.org/10.1007/s10805-025-09660-w

Mariyono, D., & Alif Hidayatullah, A. N. (2025). Navigating the moral maze: Ethical challenges and opportunities of generative chatbots in global higher education. *Applied Computational Intelligence and Soft Computing*, 2025(1), 8584141. https://doi.org/10.1155/acis/8584141

Marzal García-Quismondo, M. Á., Parra-Valero, P., & Martínez-Cardama, S. (2024). A look at critical information literacy from Europe's educability project. *The Journal of Academic Librarianship*, 50(5), 102917. https://doi.org/10.1016/j.acalib.2024.102917

Matjie, M. A., Nethavhani, A., & Matlakala, M. (2026). AI and the digital divide in education. *Frontiers in Computer Science*, 8, 1759027. https://doi.org/10.3389/fcomp.2026.1759027

Muthukrishna, M., Dai, J., Panizo Madrid, D., Sabherwal, R., Vanoppen, K., & Yao, H. (2025). AI can revolutionise education but technology is not enough: Human development meets cultural evolution.

Journal of Human Development and Capabilities, 26(3), 482–492. https://doi.org/10.1080/19452829.2025.2517740

Nopas, D. (2025). Algorithmic learning or learner autonomy? Rethinking AI's role in digital education. *Qualitative Research Journal*. https://doi.org/10.1108/QRJ-11-2024-0282

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, 10(1), 89. https://doi.org/10.1186/s13643-021-01626-4

Pham, N., Do, M. K., Dai, T. V., Hung, P. N., & Nguyen-Duc, A. (2025). FAIREDU: A multiple regression-based method for enhancing fairness in machine learning models for educational applications. *Expert Systems with Applications*, 269, 126219. https://doi.org/10.1016/j.eswa.2024.126219

Rahman, M. I. (2026). Algorithmic bias and transparency in artificial intelligence tools for nursing education: A scoping review. *Nurse Education in Practice*, *92*, 104756. https://doi.org/10.1016/j.nepr.2026.104756

Roshanaei, M. (2024). Towards best practices for mitigating artificial intelligence implicit bias in shaping diversity, inclusion and equity in higher education. *Education and Information Technologies*, 29(14), 18959–18984. https://doi.org/10.1007/s10639-024-12605-2

Sha, L., Rakovic, M., Das, A., Gasevic, D., & Chen, G. (2022). Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education. *IEEE Transactions on Learning Technologies*, 15(4), 481–492. https://doi.org/10.1109/TLT.2022.3196278

Urbina, J. T., Vu, P. D., & Nguyen, M. V. (2025). Disability ethics and education in the age of Artificial Intelligence: Identifying ability bias in ChatGPT and Gemini. *Archives of Physical Medicine and Rehabilitation*, 106(1), 14–19. https://doi.org/10.1016/j.apmr.2024.08.014

Von Winckelmann, S. L. (2023). Predictive algorithms and racial bias: A qualitative descriptive study on the perceptions of algorithm accuracy in higher education. *Information and Learning Sciences*, 124(9/10), 349–371. https://doi.org/10.1108/ILS-05-2023-0045

Wongvorachan, T., Bulut, O., Liu, J. X., & Mazzullo, E. (2024). A comparison of bias mitigation techniques for educational classification tasks using supervised machine learning. *Information*, 15(6), 326. https://doi.org/10.3390/info15060326